Electronic Letters on Computer Vision and Image Analysis 14(1):21-37, 2015

Video summarization by an innovative method in shot detection

Ali Mohammad Ahmadzade* and Hassan Farsi*

* *Department of Electrical and Computer Engineering,University of Birjand, Birjand, Iran*

Received 13th Jan 2015; accepted 14th Apr 2015

Abstract

The aim of a video summarization system is to provide a set of key frames which contain the most important parts of video. This system results in efficient storage, quick browsing, and retrieval of collection of video. In this paper, we propose a new summarization system which firstly divides the video into meaningful shots using an innovative and fast method, and then we sample the video frames belonging to each shot which results in 97% reduction in under-process frames. Then, using various characteristics of sampled frames such as color histogram, correlation and moment of inertia, wepropose anadaptive aggregation function for combination of these characteristics (differences) and extraction of key frames. The proposed systemis evaluated using 250 manual key frames constructed by human operators from 50 downloaded videos. The obtained results show that the proposed system provides better results compared to 6 different traditional methods.

Key Words: Video Summarization, Hybrid Method, Shot Detection, Key Frame, Adaptive Sampling

1 Introduction

Nowadays, a large number of video data is produced over the world and processing these data needs many sources such as time, human force and powerful hardware video data including image, sound and text. This causes that the process of video data become more complex and results in some problems. First problem is to effectively organize raw video data produced by different sources. This requires defining special frames to save video data. To solve this problem, specialists have defined special formats which have been accepted as standards by International Telecommunications Union (ITU). For example, famous formats are MPEG2 or H.263 [1].Second problem is to find effective mechanism to access video data that has been saved in view of a special standard. To overcome to this problem, appropriate indexing methods are used; there by important information of video is indexed for re–access. Video summarization is a set of video frames which includes this indexed information.

Video summarization is constructed in different forms. Two common methods are static and dynamic [2]. The static video summarization deals with key frames. The key frames are fixed video images which include the most important video contents [3].The dynamic video summarization includes a sequence of small shots which have been placed in time order. Advantage of the dynamic method is that it keeps dynamically the video contents while key frame extraction has more flexibility to show the video contents and can be used as different pre–processors [4].

Correspondence to: <hfarsi@birjand.ac.ir>

Recommended for acceptance by <Debnath Bhattacharyya>
DOI http://dx.doi.org/10.5565/rev/elcvia.697
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

Normally, the key frame is classified in two common forms based on text and content. In the first type in which the text is accomplished with the video, a desired word and or a sentence is chosen, then a frame containing that text is selected. This method is costly and time–consumer (Figure 1). The second type which is more common is performed by processors based on features extracted from a video. This method is cheaper and quicker than the first method, but due to difference of conception of human and machine, is unreliable [5]. So, in order to mitigate this difference, researchers have used different descriptors like color [6], dynamic content [7], speech [8] and object [9] for process of summarization.

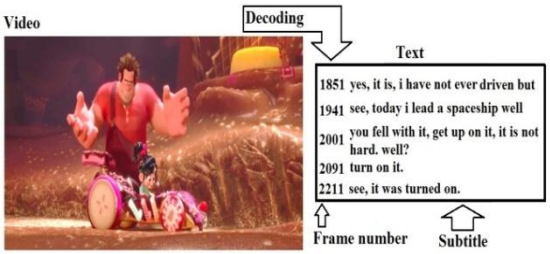

Figure1: Key frame extraction based on text.

As shown in Figure 2, the most important parts of this paper include the following issues:

1. Video Shot Boundary Detection, VSBD, by investigating features of two successive frames.
2. Process of adaptive sampling in each shot and selection of candidate frames in it.
3. Using optimized VSUKFE (Video summarization using key frame extraction) method for selection of key frames.

This paper is organized as follows: in section 2, a review of previous works is indicated in area of video summarization and in section 3, the proposed method is described. In section 4, the obtained results using the proposed method are compared to traditional methods by standard criteria and finally, conclusions are drawn in section 5.

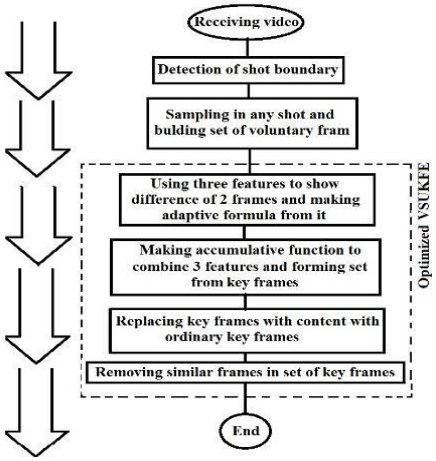

Figure 2: Flow chart of the most important parts of the proposed system.

2 A Review of Past Works

So far, various methods have been introduced for video summarization. In some of these methods, all video frames are used to select the key frames. In fact, these methods without inducing additional cost in beginning of summarization process, have tried to further simplify the system, but due to process of details on all frames, they require a lot of time. For example, VSUKFE method [6] by sampling all frames tries to decrease processed data [6]. This method, by applying an accumulated mechanism, combines visual features extracted from correlation of colorful channels, color histogram and moments of inertial to find the key frames. In this method, since the shot boundaries are considered, it is possible to select the frames of shot

boundary as key frames, due to their difference with other frames. Also, it is possible short shots and therefore some key frames are removed when sampling is performed.

The most common methods of video summarization divide video into very meaningful parts for better manipulation, and then low level features are extracted from the frames. Next, by using the extracted low level features, the key frames are indexed and finally, the indexed frames are placed side by side for video presentation. For example, in a method which uses the video summarization to produce storyboard of film, STIMO [10],to divide the video, it uses feature of HSV color space distribution in a frame and quick clustering algorithm to place similar frames in one group. Finally, it selects key frames by choosing representative frame in each group. In static video summarization method, VSUMM [11], after sampling all frames, it divides remained frames by using color descriptor, k–means clustering algorithm and Euclidean distance to select the key frames in each cluster. The advantage of based clustering methods is that redundancy and repetition are less in set of key frames compared to other methods, but the clustering algorithm is time–consuming and do not consider temporal information of key frames.

There are also other methods that use the shot as the smallest meaningful part of a video to extract a key frame. For example, in [12] and [13], the shots are firstly converted into sub–shots by clustering the shots and then, entropy of each sub–shot image is calculated. The frame having maximum entropy is considered as the key frame of that sub–shot. There are different methods to shot the video, for instance, in joint comparison method [14]which is based on histogram, the values of color histogram differences related to successive frames are obtained, then the values of abrupt and gradual shot boundaries are specified by using two pre–determined thresholds. This method is not sensitive to movement of camera and the objects inside image and it recognizes gradual shot boundaries but it is sensitive to noise and due to fixed thresholds, it is not suitable for different video types. Another common method is based on image edges [15]. In this method, the image edges are firstly detected, then change rate of the edge is computed by computing the number of incoming and outgoing the edges related to current frame. Next, the shot boundary is specified by comparing with a pre–determined threshold. This method is a little sensitive to noise but it is more sensitive to sudden movements of camera and objects inside film.

In the proposed method, we have used advantages of a fast method for shot detection and adaptive sampling. It prevents the deletion of information in short shots, and reduces the size of the required information to be processed. Also, it removes the meaningless inter-shot frames. Furthermore, during the comparison of candidate frames, we consider their contents and select the frames with more contents as key frames. This approach is based on human perception.

3 The Proposed Method

Figure 2 demonstrates the proposed method. Successive video frames are firstly received and then divided into meaningful parts of shot by identification of shot boundaries. Next, to reduce processed data in every shot, an adaptive sampling is performed and then by applying the optimized VSUKFE method on representative frames, the key frames are selected. In following, we describe all stages in details.

3.1 Video Shot Boundary Detection (VSBD)

Since final aim of the system is to summarize the video as a number of key frames, in this part it is tried to use a quick and of course having less–error methods for VSBD. The first and the simplest feature which can be used to VSBD is color feature. Abrupt shot boundaries are firstly identified by using RGB colorful space and histogram difference method that is very quick, and then, gradual shot is identified by converting each frame to HSV space and averaging its color. Since color space is sensitive to noise and colorful disorders, edge descriptor and change rate of the edges are used to authenticate the identified abrupt shot boundaries. This causes that the precision and speed to be increased. In following, we describe the abrupt and gradual shot boundary detection using the proposed method.

3.1.1 Abrupt Shot Detection

As shown in Figure 3, to detect abrupt shot boundary, two stages are used; First stage is color histogram difference method in RGB space. Since the transmission from the frame of the previous abrupt shot to the frame of the next abrupt shot occurs abruptly in abrupt shot, the color histogram difference considerably

change sat the boundary of abrupt shot compared to the other adjacent differences. The color histogram difference of a part of video 'V21' in the RGB space and the HSV space is illustrated in Figure 4 and 5, respectively, in which the considerable changes at the boundary of their abrupt shot can be observed in both RGB and HSV spaces [14]. We have used RGB space in the proposed method for calculating the color histogram difference. Note that although the frame is converted into HSV color space in second stage (i.e. gradual shot detection), since in histogram difference method in Matlab software two frames are simultaneously transferred to HSV color space and then the quantization is performed, it will cause to reduce speed of system. For instance, the shots of film'V21' are detected by using only HSV color space and ittakes63.04second but if this film is shot by using RGB and HSV color spaces, it takes 25.01 second. Therefore, RGB color space is used to detect abrupt shot.

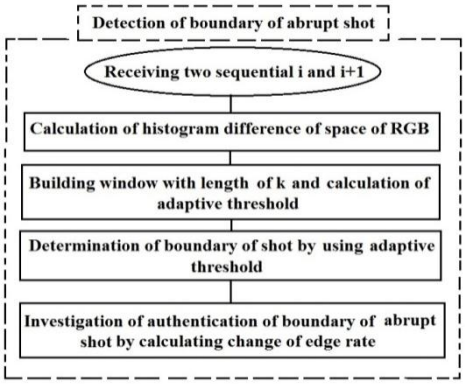

Figure 3: Flow chart of abrupt shot boundary detection.

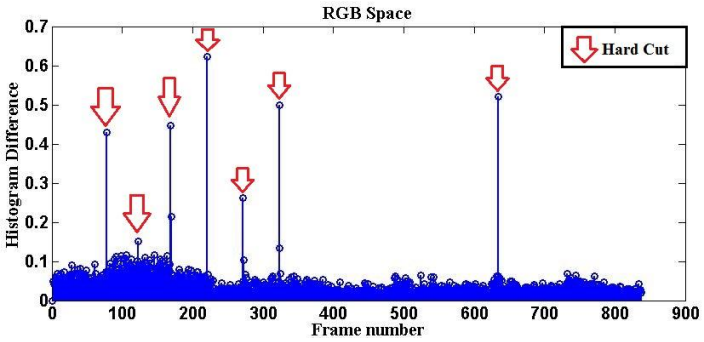

Figure 4: Color histogram difference of a part of video 'V21' in the RGB space.

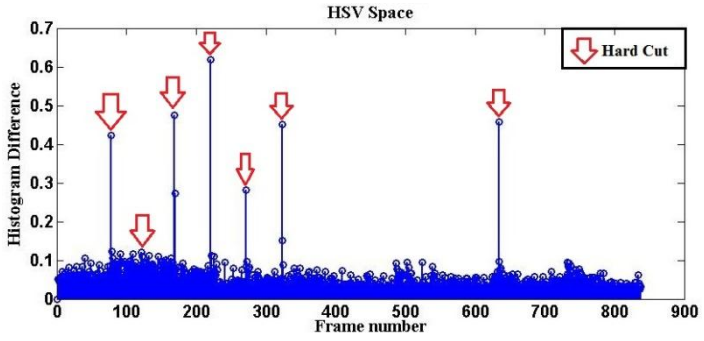

Figure 5: Color histogram difference of a part of video 'V21' in the HSV space.

After calculation of histogram difference, it is turn to detect shot boundary. In the proposed method, adaptive threshold is used as the method based on colorful features [16]. We use a moving window with 20

frames in length and threshold of TS1 is calculated as given by Equation 2. When a frame is larger than TS1, it is selected as abrupt shot.

$$\text{mean(k)}=\frac{\sum_i^{i+k} SD_i}{k} \tag{1}$$

$$\text{TS1}= \alpha \times \text{mean(k)} \tag{2}$$

Where SD_i is the histogram difference of two successive frames, k is window length and α is regulation coefficient. To determine the value of α, α is calculated for 30 abrupt shots from 30 different videos and presented in Figure 6. The value of α in Equation 2 is adopted between4 to 5 by using the minimum value of α in Figure 6.

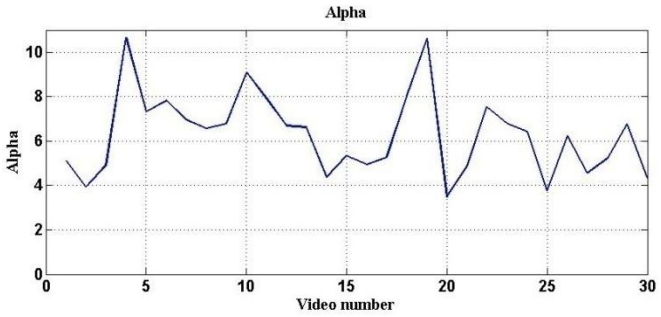

Figure 6 :Value of α for 30 abrupt shots from 30 different videos.

After initial identification of abrupt shot, we have to investigate its authentication. We use edge descriptor and change rate of edge in following steps:

i. Image edges for two successive frames are specified and sum of number of pixels belonging to image edges is calculated; edge detection is performed by Canny algorithm and the obtained image has pixel value 0 or 1.

ii. Calculation of complement of the image edges for each frame with one freedom degree of r; the edge pixels which have been set to 1 in previous stage are thickened by a square with length of r=4 (four neighbour pixels of the edge are set to 1) and then, logical NOT is driven from the obtained image (pixel 0 is converted to 1 and vice versa).

iii. Logical AND operation is applied on the obtained image in step (i) belonging to current frame and the resulted image in step (ii) related to previous frame. This causes that the previous image edges are removed and the added pixels are maintained.

iv. The obtained images show the number of incoming and outgoing edge pixels. In fact, incoming image edges are obtained by operating logical AND between current frame in step (i) with previous frame in step (ii) and outgoing image edges are calculated by operating logical AND between the previous frame in step (i) and the current frame in step (ii).

v. Calculation of ECR given by Equation 3 and comparison to a threshold.

$$ECR_n= \max\left(\frac{X_n^{in}}{\delta_n}, \frac{X_{n-1}^{out}}{\delta_{n-1}}\right) \tag{3}$$

Where δ_n indicates the number of edge pixels in nth frame, X_n^{in} and X_{n-1}^{out} are the number of incoming and outgoing edge pixels belonging to n^{th} and $(n-1)^{th}$ frame, respectively.

If the value of ECR exceeds the threshold TS3 which is adopted between 0.3 and 0.5, then the frame is selected as abrupt shot boundary.

3.1.2 Gradual Shot Detection (Inventive Method)

To detect shot boundary, total average of color of each frame is used and since HSV color space shows gradual shots better than RGB color space, the frames are firstly converted into HSV color space and then averaging is applied on its three dimensions. Figure 7 and 8 show difference of two spaces for a part of film,V21.

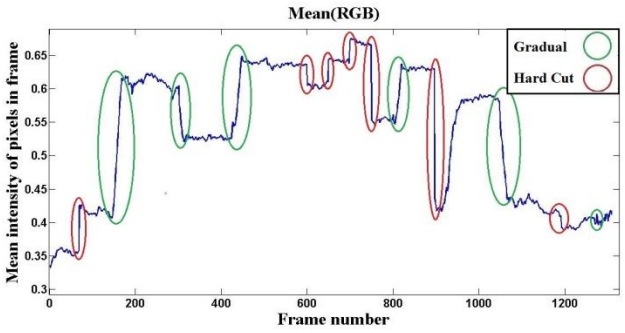

Figure 7: Average color of video frames in RGB color space.

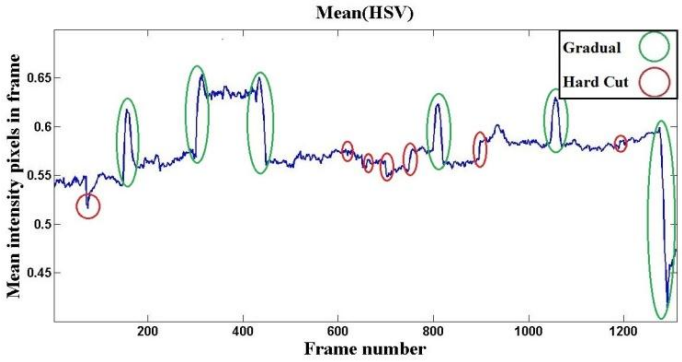

Figure 8: Average color of video frames in HSV color space.

As observed, the HSV color space indicates gradual shots by presenting a peak or deep better than RGB color space but it is unable to show abrupt shot boundaries as good as RGB color space.

Next problem is how to detect these peak or deep points in diagram of average color of HSV. To solve this problem, an inventive method based on variable steep is used as follows:

After calculation of average color of a frame, its value is saved as variable of SD1. To identify peak or deep points, a moving window is used with 21 frames in length and the value of steep is calculated by:

$$\text{Ra}=\frac{\max(Wd)-\min(Wd)}{21} \tag{4}$$

Where Wd is moving window and its values are the average color of current frame and its 20 previous frames and Ra is window steep.

The value of Ra inside variable SSD having an equal length to total number of frames is saved in the same place of window in SD1. For example, SSD curve for Figure 8 is drawn in Figure 9.

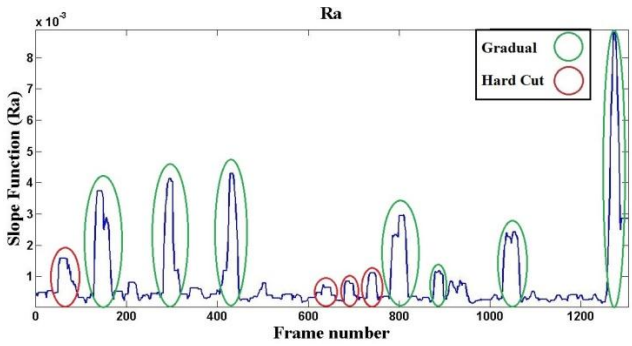

Figure 9: Variable steep belonging to Figure 8.

As seen in Figure 9, calculation of steep in a moving window causes that the peaks and deeps representing the gradual shots, are clearly observed and identified (places signed by green color). Although some points related to abrupt shots (places signed by red) are also considered as gradual shots, this problem will be solved by overlapping boundary points as described in part of division of video (sub section C). To identify these points, we use an adaptive threshold and a moving window given by:

$$\text{meanRa(k)}=\frac{\sum_{i}^{i+k} Ra_i}{k} \tag{5}$$

$$\text{TS2}=\beta\times\text{mean(k)} \tag{6}$$

Where Ra_i is window steep in frame i, k is the length of window (k=20), and β is regulation coefficient adopted between 2 to 3. To provide better detection, this window is quantized by TS2 and Equation 7.

$$\text{Wini} = \begin{cases} 0 & \text{if Ra} > Ts2 \\ 1 & \text{if Ra} < TS2 \end{cases} \tag{7}$$

If sum of μ, the last number of this quantized window, is equal to μ, then the first maximum point is chosen up to 30 frames after it as first gradual shot boundary and the first number after a peak point whose value is a half of peak value, is selected as the end of shot boundary (μ approximately is chosen between 4 to 8 by tasting various). For example, Figure 10 shows a gradual shot in film,V21, which has been quantized and identified by Equation 7 and μ=5. This method approximately identifies shot boundary. To determine the value of β, assuming μ=4, β is calculated for 30 gradual shots from 30 different videos and presented in Figure 11. The value of β in Equation 6 is adopted between 2 to 3 by using minimum value of β in Figure 11.

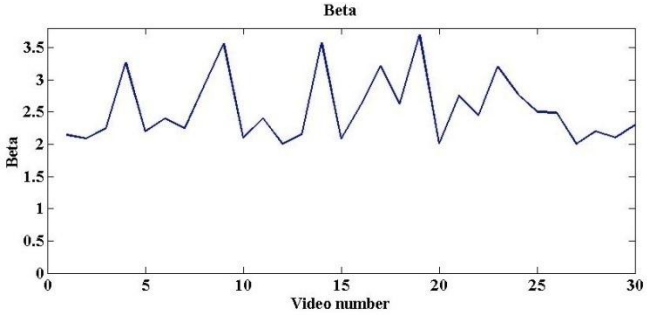

Figure 11: Value of β for 30 gradual shots from 30 different videos (μ=4).

3.1.3 Division of Video

The aim of this part is division of video by using identified shot boundaries. All identified shot boundaries are saved and compared to gradual shots. This causes that the meaningless frames representing gradual shot are removed and the frames of each shot are saved inside a separate set for sampling operation.

3.2 Adaptive Sampling

This stage is used as a pre–processing to reduce processed information and improve the performance of system. Adaptive sampling is used to remove noise frames and preserve short shots. Therefore, after specifying the shots, the average is obtained using differences of binary RGB color space histogram in each shot and the frames which are less than the obtained average are chosen as initial samples. Then, a set of candidate frames is selected by using pre–sampling method provided in VSUKFE method [6] on initial samples. Histogram difference upper than average of two successive frames is due to existing abrupt shot, gradual shot, noises of film and sudden changes inside film. For example, Figure 12 shows the differences of binary histogram belonging to a documentary film with value of its total average demonstrated by red line.

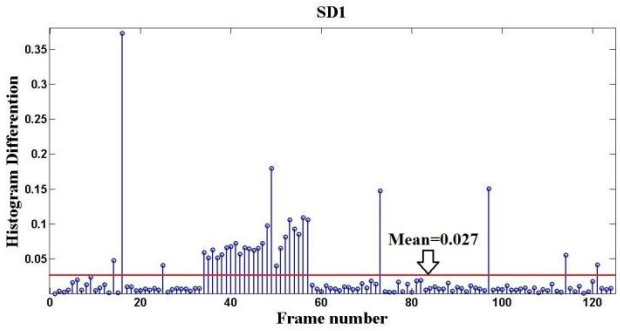

Figure 12: Histogram difference of RGB with average value.

Figure 13: Frames related to the specified points in Figure 12.

The related images to the points specified in Figure 12 are shown in Figure 13. As observed, the related points to 1 and 5 due to noise of image and the points of 2, 3 and 4 because of changing the shot, the value of their difference has exceed than average. Thus, the number of candidate frames, n_{CF}, and the set of candidate frames, CF_T, are defined as:

$$n_{CF} = \frac{n_{SF} - n_{SFM}}{\lambda}$$
$$CF_T = \{F(t+j) \mid j = 0,2\lambda,\ldots,j < \lambda \times n_{CF}\} \quad (8)$$

Where n_{SF} is the number of shot frames, n_{SFM} is the number of frames higher than histogram average, λ is the distance between two candidate frames in VSUKFE method [6] and F(t) is a frame lower than the average.

From this stage thereafter, selection of key frame is performed on set of candidate frames, CF_T.

3.3 Key Frame Selection

We have used the optimized VSUKFE method to choose key frame in any shot. In this method, like VSUKFE method [6], three criteria including inter–frame correlation between colorful channels of RGB, color histogram, and moments of inertial for obtaining displayable difference of frame are used. The frames are divided in to non–overlapped parts and then, similar parts in current frame and last key frame are compared. These criteria are briefly described as follows.

3.1.1 Correlation Frame Difference Measure

Inter–frame correlation measures the similarity between two frames based on colorful content. The correlation coefficients have been widely used to indicate the similarity between two frames. The correlation coefficients are calculated for each color channel (red, green and blue) between each section of the frames which has to be compared. Let F(t) and F(t+1) be two frames for the calculation of correlation from the set of candidate frames CFT. It is assumed that the dimension of each frame is m×n and each frame has been divided into a total of Ts sections of p×q each. Then the correlation coefficient for a section 's' for a color channel 'c' is given by:

$$r(F(t),F(t+1))s,c=\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(F_s(t)_{c,i,j}-\overline{F_{C,S(t)}})((F_s(t+1)_{c,i,j}-\overline{F_{C,S(t+1)}})}{\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q}(F_s(t)_{c,i,j}-\overline{F_{C,S(t)}})^2\sum_{i=1}^{p}\sum_{j=1}^{q}(F_s(t+1)_{c,i,j}-\overline{F_{C,S(t+1)}})^2}} \quad (9)$$

Where $F_s(t)_{c,i,j}$ is the pixel value of 'c' color channel of F(t) at row 'i' and column 'j' in section 's', and $\overline{F_{C,S}(t)}$ is the mean values of pixel values of color channel 'c' of frame F(t). The correlation coefficient is computed for each section (s=1...Ts) and color channel (c= red, green, blue). The mean of correlation of all sections is then taken to compute the overall correlation for a color channel:

$$r(F(t),F(t+1))c=\frac{1}{T_s}\sum_{k=1}^{T_s} r(F(t),F(t+1))_{k,c} \quad (10)$$

Finally, the obtained values for all color channels are combined using the mean function to obtain the result of correlation comparison measure as:

$$\rho(F(t),F(t+1))=\frac{r(F(t),F(t+1))_{red}+r(F(t),F(t+1))_{green}+r(F(t),F(t+1))_{blue}}{3} \quad (11)$$

3.1.2 Histogram Frame Difference Measure

Colorful histograms are used to measure differences based on color and due to their simplicity and power in recognition of small changes in content of color and low sensitivity to movement of objects inside frame, they have been selected. Using two different colorful spaces to measure histogram difference and correlation of two frames will show calculations of color difference from two different views. In this method, after obtaining a color histogram in RGB space, a color quantization step is applied to reduce the size of the color histogram. The quantization of the color histogram is set to 16 bins for hue components, and 8 bins for each of the saturation and intensity components. Histogram frame difference is measured by:

$$H(F(t),F(t+1))=\frac{1}{T_s}\sum_{p=1}^{T_s} 1-\sum_{l=1}^{32}\min(H_{t,p}(l),H_{t+1,p}(l)) \quad (12)$$

Where $H_{t,p}(l)$ and $H_{t+1,p}(l)$ are the color histograms of the pth section F(t) and F(t+1) respectively.

3.1.3 Moments of Inertia Difference

The moments of inertial have been used as an additional criterion to provide an imagination for image description. In VSUKFE method [6], the three moments inertia (mean, variance, skewness) are calculated by:

$$\overline{F(t)_{s,c}}=\frac{1}{p\times q}\sum_{i=1}^{p}\sum_{j=1}^{q}F(t)_{i,j}$$

$$\sigma^2F(t)_{s,c}=\frac{1}{p\times q}\sum_{i=1}^{p}\sum_{j=1}^{q}(F(t)_{i,j}-\overline{F(t)_{s,c}})^2$$

$$\gamma(F(t))_{s,c}=\frac{1}{p\times q}\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(F(t)_{i,j}-\overline{F(t)_{s,c}})^3}{(\sigma^2F(t)_{s,c})^{3/2}} \quad (13)$$

Where $\overline{F(t)_{s,c}}$, $\sigma^2F(t)_{s,c}$ and $\gamma(F(t))_{s,c}$ are the mean, variance and skewness values of color channel 'c' in section 's', respectively. Finally, these values are combined to form a moments of inertia feature vector μ_t of frame F(t). The size of vector is 9×Ts (3 moments for each color channel (red, green and blue)). The moments of inertia difference measure between to frames F(t) and F(t+1) is computed by using Euclidean distance between the respective feature vectors as:

$$\mu(F(t),F(t+1))=\sqrt{\sum_{i=1}^{9T_s}(\mu_t(i)-\mu_{t+1}(i))^2} \quad (14)$$

An accumulative function by using subsidiary values of features difference is constructed for comparison with a threshold. According to VSUKFE method, when the value of accumulative function is less than main threshold, τ, high change in content of frame will be unknown compared to key frame and so it is removed. This seems right but it causes to remove the frames with lower accumulative function and with more content in shot. For example, Figure 14 shows 8 frames which are mutually belonging to one shot. Since the frames 3, 5, 7 and 9 have been placed at beginning of the shot, they have been chosen as key frames and the frames 4, 6, 8 and 10 are removed due to their accumulative function values indicated in Table 1 and finally, the frames in beginning of the shot in spite of having less importance than next frames in the same shot, are selected as key frames.

Figure 14: Frames with lower accumulative function, but with more contents.

Frames	3,4	5,6	7,8	9,10
$d_{\rho H\mu}$	-0.70	0.88	-0.44	0.83

Table 1: Accumulative function values related to frames in Figure 14.

In order to overcome to this problem, entropy of image and HSV color space are used due to adjacency to conception of human and color. This is achieved as follows; in one shot, when the accumulative function value of key frame is less than the value one in next frame, the entropy of current frame is compared to the entropy of key frame and if it is higher than the one in key frame then the frame is considered as key frame. This process continues till selecting a new key frame. When a new key frame is identified, the previous key

frame is replaced by the frame whose entropy is more than the key frame one, due to having more colorful contents. This means that the previous key frames are replaced by the frames having more contents. For example, entropy values of the frames in Figure 14 are shown in Table 2. This indicates that the previous key frames are replaced by the frames 4, 6, 8 and 10 during the process of selecting the key frames.

Number Frame	Fr.3	Fr.4	Fr.5	Fr.6	Fr.7	Fr.8	Fr.9	Fr.10
Entropy	5.317	5.502	5.584	5.688	5.658	5.937	6.526	6.537

Table 2: entropy values for Figure 14.

3.4 Removing Similar Frames

After choosing the key frame in each shot, it requires to remove repetitive key frames having equal contents. This stage is performed as same as the method reported in VSUMM. In this method, Euclidian distance between colorful histogram of HSV color space belonging to the selected key frames is obtained and the key frames having similarity more than 50% are considered to be similar to each other and one of them is removed. This process is performed for all the remained key frames and finally a set of key frames are constructed at the end.

4 Experiment and Result

According to the proposed methods in section 3, assessment of the proposed system includes 3 stages. In first stage, the proposed video division, or VSBD system, is compared against the existed VSBD methods and then, authentication of the proposed sampling system will be examined on key frames and finally, the proposed video summarization system is compared to several methods.

4.1 VSBD System

To assess VSBD system, we use the criteria reported in [17] which are named Recall and Precision and given by Equation 15.

$$\text{recall}=\frac{N_c}{N_c+N_m}\times 100\%$$
$$\text{precision}=\frac{N_c}{N_c+N_f}\times 100\% \quad (15)$$

Where $N_c$is the number of detected shots correctly, N_m is the number of undetected shots and N_f is the number of false–detected shots. The experiments are performed on 20 videos downloaded from [20] and 10 videos downloaded from [21].Totally, 307 abrupt shot and 194 gradual shot are investigated by the proposed system.The second source has been selected due to updating its videos. Three videos are used for comparison of the proposed method against the reported method [16]. The results have been obtained by MATLAB 2013b software and using a computer system with four–core processors of 2.2GHz and 8 GB RAM. For example, Table 3 shows specifications of 10 tested films (n_{AF} is total number of video frames , n_{ASG} is total number of gradual shots and n_{ASH} is total number of abrupt shot boundaries). Simulation results are shown using the proposed method in Table 4. In addition, the evaluation results are presented in Table 5 using the obtained data in Table 4.

Number	Films	n_{ASH}	n_{ASG}	n_{AF}
1	V21	19	10	3283
2	V24	0	11	1815
3	V60	10	2	2092
4	V36	9	11	4565
5	NASA	4	3	4746
6	Ferdosipooer	4	0	957
7	Help cow	15	1	1836
8	Anni 007	5	5	1590
9	UGS01-002	11	3	2766
10	UGS01-011	2	15	3606

Table 3: Specifications of videos in stage of VSB Devaluation.

Number of film	Gradual shot			abrupt shot		
	N_f	N_m	N_c	N_f	N_m	N_c
1	18	1	0	7	1	1
2	-	-	0	8	1	1
3	8	1	1	2	0	-
4	8	1	0	9	2	0
5	4	0	0	3	0	0
6	4	0	0	-	-	-
7	14	1	1	1	0	0
8	5	0	0	5	0	1
9	10	1	0	3	0	1
10	2	0	0	14	2	1

Table 4: Detected shots using the proposed method.

Gradual shot Boundary		Abrupt shot Boundary	
Precision	Recall	Precision	Recall
91 %	89 %	98 %	95 %

Table 5: The evaluation results using the obtained data in Table. 4.

Table 6 compares the obtained results by the proposed method against the reported method in [16] using three common videos in both methods which are videos of 8, 9 and 10.

Type of method	Abrupt shot boundary		Gradual shot boundary	
	Recall	Precision	Recall	Precision
Method [16]	88 %	100 %	82 %	79 %
Proposed method	94 %	100 %	91 %	88 %

Table 6: Comparison of the proposed VSBD method against the reported method in [16].

As observed in Table 6,the proposed method provides better results than the method reported in [16].

Finally, the obtained results for the shot detection system on 30 videos are shown in Table 7.

Gradual shot Boundary		Abrupt shot Boundary	
Precision	Recall	Precision	Recall
91 %	94 %	83 %	87 %

Table 7: The evaluation results for the proposedshot detection system.

Since the proposed VSBD method recognizes the video shots precisely and removes the frames between the shots for next stage, meaningless inter-shot frames will not be displayed in key frames. For instance, in assessed video films shown in Figure 15, inter-shot frames exists in set of key frames using the methods reported in DT[18], STIMO [10], and VSUMM [11] while these frames are removed by the proposed VSBD system.

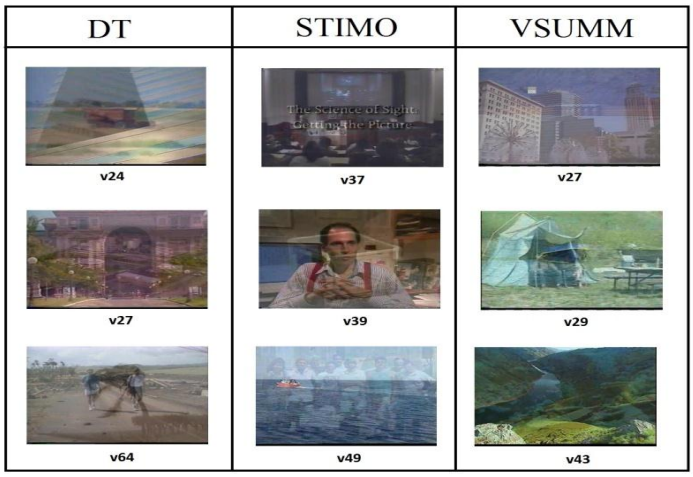

Figure 15: Existence of inter-shot frames in key frame using other methods.

4.2 Adaptive Sampling System

To assess sampling system in summarization process, its direct effect is investigated on choosing key frames, or in other words, this question will be numerically answered whether the proposed sampling system is able to remove key frames. We test 10 videos and the percentage of authentication of sampling system, given by Equation 16, is calculated.

$$\text{Authentication of sampling system} = \frac{n_{PSR}}{n_{US}} \tag{16}$$

Where n_{US} is the number of key frames chosen by human user or manually and n_{PSR} is the number of key frames removed by the sampling system. In Table 8, 3 samples of investigated videos are shown by considering total number of voluntary frames in video, n_{ACF}, total number of frames of video, n_{AF}, and total number of video shots, n_{AS}.

Name	n_{PSR}	n_{US}	n_{ACF}	n_{AF}	n_{AS}
V44	11	2429	62	12	0
V50	8	4825	119	9	0
V63	11	2308	63	8	1

Table 8: A sample of investigated videos using the proposed sampling system.

In this experiment, by investigating on 30 videos with 87791 frames, 2111 frames have been chosen as candidate frames and among of 906 key frames that have been manually chosen by human user from these videos, there are not only four frames inset of candidate frames. This means that the authentication of sampling method in summarization process is roughly 99%. The obtained results show that the sampling process reduces processed data in stage of summarization around 98% and causes to increase speed and performance of system.

4.3 Video Summarization

To assess the proposed summarization system, we use the criteria reported in VSUMM [11]. In this plan of assessment, the summarized video frames are manually constructed by user and used as a reference to compare to those obtained by automatic methods. Each key frame obtained by automatic method is compared to manual ones based on Manhattan distance between color channel histograms of two frames in HSV color space. Two key frames are similar if the distance is equal to $\tau_{com} = 0.5$, as suggested in [11]. Also, in this assessment method, the quality of the obtained key frames using the automatic methods is evaluated by two criteria; precision rate, CUS_{A}, and error rate, CUS_E, given by:

$$CUS_{\mathrm{A}} = \frac{n_{\mathrm{mAS}}}{n_{\mathrm{US}}}$$

$$CUS_E = \frac{n_{\mathrm{m'AS}}}{n_{\mathrm{US}}} \tag{17}$$

Where n_{mAS} is the number of key frames which have been obtained by automatic summarization (AS), $n_{\mathrm{m'AS}}$ is the number of key frames, not provided from automatic summarization and n_{US} is the number of manual key frames constructed by user. The proposed system has been applied on 50 videos chosen from [20]. This set has been constructed by Avila in which, there are summaries made by human users. These videos are in format of MPEG–1 and with time length between 1 to 4 min (30 frames/sec, 352×240 pixels) and contain different subjects such as documentary, educational, historical and conference. The manual summaries have been constructed by 50 people. For summarization, 5 videos have been given to each user and so there are250 manual summarized frames (key frames)for comparison. In addition to manual summaries and VSUMM method, the summaries obtained by other methods such as OV[19], DT[18] and STIMO [10]are accessible in [22]. Therefore, we compare the obtained results by the proposed method to those obtained by OV, DT, STIMO and VSUMM methods.

For evaluation of the proposed system, the summaries obtained by the proposed system for each video are compared to 5 related manual summaries. Totally, 176,000 video frames are investigated by the proposed system and 2265 frames are chosen as key frames and compared to 2162 manually–made frames. This indicates that 1898 frames are correctly recognized. Therefore, the values of CUS_{A} and CUS_{E} are 85.6% and 20.3%, respectively. The obtained results are shown in Table 9.

Video summarization methods	CUS_{A}	CUS_{E}
OV	0.70	0.57
DT	0.53	0.29
STIMO	0.72	0.58
VSUMM	0.85	0.38
VSUMM2	0.70	0.27
VSUKFE	0.80	0.32
Proposed method	0.85	0.20

Table 9: Comparison of the proposed video summarization method with other methods.

As an example, Figure 16 shows the manual key frames by 5 users for video, V60. By applying the different video summarization systems including the proposed method on video V60, the obtained key frames are shown in Figure 17. Compared to the reference key frames shown in Figure 16 and by considering the values of CUS_{A} and CUS_{E}, we could say that the proposed system results in lower error rate and higher accuracy than other methods.

5 Conclusion

In this paper, we proposed a new method for video summarization based on a powerful key frame selection algorithm. In the proposed method, to manipulate video information and remove meaningless frames, the video content is divided into meaningful parts of shot.

VSBD operation is performed in two stages. Firstly, abrupt shot boundaries are identified by using histogram difference, and then, their authentication is investigated by calculating edge rate. Finally, gradual shot boundaries are specified by using an inventive method. In VSBD stage, the adaptive threshold and regulation constants are used to reduce dependency of VSBD system to the video contents. After applying the adaptive sampling and constructing a set of candidate frames, the optimized VSUKFE method is used for key frame selection. To evaluate the proposed system, a criterion introduced by Avila is used. The obtained results show that the proposed video summarization system compared to 6 different systems provides higher accuracy for similar video sources. Since the proposed system is constructed by different algorithms such as VSBD and sampling, the performance of these algorithms are individually evaluated and they have provided acceptable results.

In the proposed method, in spite of traditional methods which process all frames in a video to select the key frames, we have taken advantages of a fast method for shot detection and adaptive sampling. The proposed method prevents the deletion of information in short shots, and reduces the size of the information to be processed up to 98%. Also, it removes the meaningless inter-shot frames that may be selected as key frames because of their differences with the other frames. Furthermore, in spite of VSUKFE[6] method, during the comparison of candidate frame we consider their contents and select the frames with more contents as key frame. This approach is based on to human perception.

Figure 16: Manualsummaries (key frames) provided by human user of videoV60.

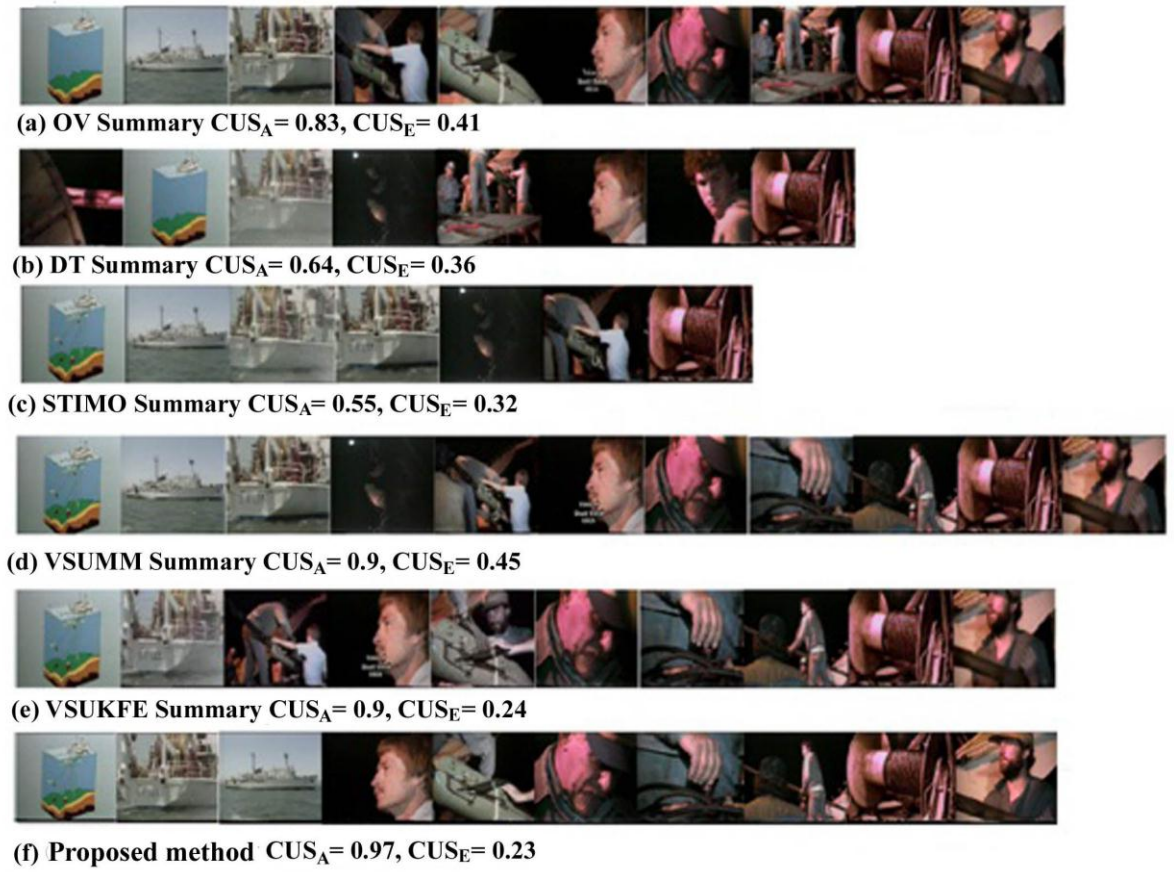

Figure 17: Comparison of the proposed method to other methods for videoV60.

References

[1] A. Luther, A. Inglis, *Video Engineering*, United States, McGraw-Hill Education, 1999.

[2] Y. Li, T. Zhang, D. Tretter, "An overview of video abstraction techniques", *Technical Report HP Laboratory*, HP-2001-191, 2001.

[3] G. Ciocca, R. Schettini, "Innovative algorithm for key frame extraction in video summarization", *Journal of Real-Time Image Processing, Springer*, Vol: 1, NO: 1, pp: 69-88, 2006. <http://dx.doi.org/10.1007/s11554-012-0278-1. >

[4] H.H. Kim, Y.H. Kim, "Toward a conceptual framework of key-frame extraction and storyboard display for video summarization", *Journal of the American Society for Information Science and Technology*, Vol: 61, NO: 5, pp: 927–939, 2010. < http://dx.doi.org/ 10.1002/asi.21317>

[5] A. Massimiliano, *Extracting and summarizing information from large data repositories*, Ph.D. Thesis, University of Naples Federico II, Italia, November 2006.< http://dx.doi.org/ 10.6092/UNINA/FEDOA/577>

[6] N. Ejaz, T.B. Tariq, S.W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism", *Visual Communication Image Representation, Elsevier*, Vol: 23, pp: 1031–1040, June 2012. <http://dx.doi.org/10.1016/j.jvcir.2012.06.013>

[7] F. Chen, M. Cooper, J. Adcock, "Video Summarization Preserving Dynamic Content", *the International Workshop on TRECVID Video Summarization*, 2007. <http://dx.doi.org/10.1145/1290031.1290038>

[8] C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, E.J. Delp, "Automated video program summarization using speech transcripts", *IEEE Transactions on Multimedia*, Vol: 8, pp: 775-791, 2006. <http://dx.doi.org/10.1109/TMM.2006.876282>

[9] A.M. Ferman, B. Gunsel, A.M. Tekalp, "Object-Based Indexing of MPEG-4 Compressed Video", *Symposium IS&T/SPIE. on Electronic Imaging*, 1997.

[10] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, " STIMO: STIll and moving video storyboard for the web scenario", *Multimedia Tools and Applications, Springer*, Vol: 46, NO: 1,pp: 47–69, 2009. <http://dx.doi.org/10.1007/s11042-009-0307-7>

[11] S.E.D. Avila, A.B.P. Lopes, L.J. Antonio, A.d.A. Araujo, "VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method", *Pattern Recognition Letters, ELSEVIER*, Vol: 32, NO: 1, pp: 56–68, 2011. <http://dx.doi.org/10.1016/j.patrec.2010.08.004>

[12] R. Pan, Y. Tian, Z. Wang, "Key-frame Extraction Based on Clustering", *Supported by Natural Science Basic Research Plan in China, IEEE*, 2010. <http://dx.doi.org/10.1109/PIC.2010.5687901>

[13] L. Pan, X. Wu, X. Shu, "Key Frame Extraction Based on Sub-shot Segmentation and Entropy Computing", *Jiangnan University, China, IEEE*, 2009. <http://dx.doi.org/10.1109/CCPR.2009.5343990>

[14] H.J. Zhang, A. Kankanhalli, S.W. Smoliar, "Automatic partitioning of full-motion video", *Multimedia Systems, Springer*, 1993.

[15] A. Jacobs, A. Miene, G.T. Ioannidis, O. Herzog, "Automatic shot boundary detection combining color, edge, and motion features of adjacent frames", *Center for Computing Technologies, University of Bremen*, 2004.

[16] H. Zhang, R. Hu, L. Song, "A Shot Boundary Detection Method Based on Color Feature", *International Conference on Computer Science and Network Technology, Wuhan University ,China, IEEE*, 2011. <http://dx.doi.org/10.1109/ICCSNT.2011.6182487>

[17] G. Boccignone, A. Chianese, V. Moscato, , A. Picariello, "Foveated Shot Detection for Video Segmentation", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol: 15, NO: 3, pp: 365-377, 2005. <http://dx.doi.org/10.1109/TCSVT.2004.842603>

[18] P. Mundur, Y. Rao, Y. Yesha, "Keyframe-based video summarization using Delaunay clustering", *International journal Digital Library*, Vol: 6, NO: 2, pp: 219–232, 2006. <http://dx.doi.org/10.1007/s00799-005-0129-9>

[19] D. DeMenthon, V. Kobla, D. Doermann, "Video summarization by curve simplification", *6thACM International Conference on Multimedia. NY, USA*, pp: 211–218, 1998. <http://dx.doi.org/10.1145/290747.290773>

[20] Available online at website: www.open–video.org.

[21] Available on line at website: www.aparat.com.

[22] Available online at website: www.npdi.dcc.ufmg.br/VSUMM.